Falcon-H1R Performance on GPQA Diamond and Advanced Reasoning Benchmarks

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: Which frontier language models achieve highest scores on GPQA Diamond Humanity Last Exam and difficult reasoning benchmarks v20. 6 claims were extracted from source literature; 5 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 8.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Falcon-H1R: Pushing the Reasoning Frontiers with a Hybrid Model for Efficient Test-Time Scaling. Research question: Which frontier language models achieve highest scores on GPQA Diamond Humanity Last Exam and difficult reasoning benchmarks v20.

2 Methodology

Systematic literature search across multiple databases yielded 4 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.2/10.

3 Results

4 papers retrieved. 6 claims extracted; 5 independently verified. Quality review score: 8.2/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
Falcon-H1R is a 7B-parameter reasoning-optimized model.	✓	0.28
Falcon-H1R matches or outperforms SOTA reasoning models that are 2x to 7x larger across a variety of reasoning-intensive	✓	0.19
Falcon-H1R utilizes a hybrid-parallel architecture design to achieve faster inference.	✓	0.17
Falcon-H1R leverages the DeepConf approach to achieve state-of-the-art test-time scaling efficiency.	✓	0.26
Falcon-H1R offers substantial improvements in both accuracy and computational cost compared to previous methods.	×	0.14
Falcon-H1R was trained using efficient SFT and RL scaling strategies.	✓	0.16

References

- https://doi.org/10.48550/arxiv.2501.17805
- https://openalex.org/W7161353827
- https://openalex.org/W7119234045